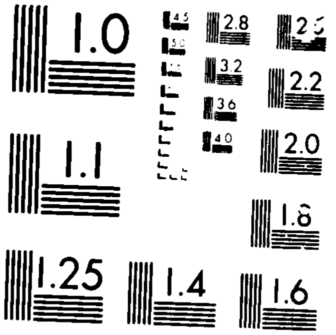
AD-A168 921 DESIGN AND ANALYSIS OF EXPERIMENTS AND MORE REALISTIC 1/1
TECHNIQUES FOR DATA ANALYSIS(U) PRINCETON UNIV NJ
J W TUKEY MAR 86 ARO-19442.26-MA ARO-14244.15-MA
UNCLASSIFIED DAAG29-79-C-0205 F/G 12/1 NL



MICROCOP' CHART

2

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO 14244-15-MA ARO 16669.28-MA ARO 19442.26-MA	2. GOVT ACCESSION NO. N/A	3 RECIPIENT'S CATALOG NUMBER N/A
TITLE (and Subtitle) Design and Analysis of Experiments and More Realistic Techniques for Data Analysis		5 TYPE OF REPORT & PERIOD COVERED Final Report 30 Jun 76 - 31 Jan 86
		6 PERFORMING ORG. REPORT NUMBER
AUTHOR(s) John W. Tukey		8 CONTRACT OR GRANT NUMBER(s) DAAG29-76-G-0298 DAAG29-79-C-0205 DAAG29-82-K-0178
PERFORMING ORGANIZATION NAME AND ADDRESS Princeton University Princeton, NJ 08540		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE March 1986
		13. NUMBER OF PAGES 44
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Graphic Analysis	Analysis of Experiments
Sampling	Data Analysis
Randomized Experiment	Confidence Intervals
Design of Experiments	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The research involved has been mainly focused on "procedures" - - on things to do with date - - on data analysis techniques.

Some of the more innovative contributions include:

graphical analysis of variance, configural polysampling, so far the only direct approach to robustness in actual finite-sized samples, compromise MLEs as a more flexible, often closely approximate, approach to

20. ABSTRACT CONTINUED

robustness, calculations for many explicit randomizations as the trustworthy and stringent analaysis for randomized experiment, simple devices for presenting motion using an overhead projector, work on pushback techniques and related ideas, effective confidence intervals based on only 2 of the observed values in a sample, new insight into multiple comparisons, fits with many exact-zero residuals sharing other good properties, deeper insight into, and robust procedures for, "analysis of variance."

ARO 14244.15-MA
ARO 16669.28-MA
ARO 19442.26-MA

Design and Analysis of Experiments *and* More Realistic Techniques for Data Analysis

by

John W. Tukey

FINAL REPORT

1976 - 1986

supported by the

U. S. Army Research Office (Durahm)

Executive Summary

The pages that follow sketch the content of nearly 100 publications (99 issued or scheduled), of which 12 are books. It also covers more than 50 technical reports, and 13 Ph.D. theses, as well as considerable amounts of unreported and unpublished material.

The work involved has been mainly focused on "procedures" - - on things to do with data - - on data analysis techniques.

Some of the more innovative contributions include:

1) Exploratory graphical analysis of variance.

2) Configural polysampling, so far the only direct approach to robustness in actual finite-sized samples.

3) Compromise MLEs as a more flexible, often closely approximate, approach to robustness.

4) Calculations for many explicit randomizations as the trustworthy and stringent analysis for randomized experiment.

5) Simple devices for presenting motion using an overhead projector.

6) Work on pushback techniques and related ideas.

7) Effective confidence intervals based on only 2 of the observed values in a sample.

8) New insight into multiple comparisons.

9) Fits with many exact-zero residuals sharing other good properties.

10) Deeper insight into, and robust procedures for, "analysis of variance".

John W Tukey

John W. Tukey
Princeton, 18 March 1986

Final Report on work at Princeton University on "Design and Analysis of Experiments," and "More Realistic Techniques for Data Analysis," supported by the U. S. Army Research Office (Durham).

Principal Investigator

John W. Tukey

Co-investigators

Henry I. Braun

Contents:

Attached: Technical detail for July 1985—January 1986.

1. Personnel 1974 — 1986

Faculty

Henry L Braun 1974-79
Colin Goodall 1985-86
Elvezio Ronchetti 1983-86
Andrew F. Siegel 1979-83
Gary Simon 1974
John W. Tukey 1974-86

Visiting Faculty (some short term)

Alfio Marazzi 1978S
James McBride 1979
Henry Wynn 1978S
Allan Seheult 1981S
Michael Cohen 1981-82
Paul Velleman (off site) 1982-83

Post-Doctoral

Daryl Pregibon 1979-80

Graduate Students

Dhammika Amaratunga 1982-84 (Ph.D. 1984)
Katherine Bell (later Krystinik) 1980-81 (Ph.D. 1981)
David Coleman 1978
George S. Easton 1984-85 (Ph.D. 1985)
Steven Finch 1974 (Ph.D. 1974)
JoAnne Goldberg 1979-80
Katherine Hansen 1984-85 (still a graduate student)
Paul S. Horn 1980-81 (Ph.D. 1981)
Clifford Hurvich 1980-81 (Ph.D. 1985)
Eugene Johnson 1984-85 (Ph.D. expected 1986)
Karen Kafadar 1978-79 (Ph.D. 1979)
Lois Kellerman 1974-75
David A. Lax 1975 (Ph.D. Harvard 1981
Stephan Morgenthaler 1981-83 (Ph.D. 1983)
Ha Nguyen 1984-86 (Ph.D. probable 1986)
Fanny (Zambuto) O'Brien 1983-84 (Ph.D. 1984)
David Pasta 1974 (graduate study continued at Stanford 1974-86)
Lincoln Polissar 1974 (Ph.D. 1974)
David Rubin 1983-1984 (still a graduate student)
Michael D. Schwarzschild 1974-79 (Ph.D. 1979)
P. Slasor 1984-85 (still a graduate student)
David E. Tyler 1974-75 (Ph.D. 1979)
Paul Velleman 1973-1974 (Ph.D. 1975)

Research Assistant

E. Olszewski 1983-86

Undergraduate students

Andrew Bruce 1980
George Grover 1982
Alan Minkoff 1979-80

2. Publications 1974-1986

Books

Breckenridge, Mary (1983). *Age, Time and Fertility: Applications of Exploratory Data Analysis,* Academic Press. (cp. Technical Report 143).

Ronchetti, Elvezio. M., (and Hampel, F. R., Rousseeuw, P. J. and Stahel, W. A., as co-authors) (1986d). *Robust Statistics: The Approach Based on Influence Functions,* John Wiley & Sons, Inc., New York.

Siegel, Andrew, F., (and Launer, R. L., as co-editors) (1982). *Modern Data Analysis,* Academic Press, New York.

Siegel, Andrew, F., (and Romano, J. P., as co-author) (1986). *Counterexamples in Probability and Statistics,* Wadsworth, Inc., Belmont CA.

Tukey, John W. (1977a). *Exploratory Data Analysis, First Edition,* Addison-Wesley Publishing Company, Reading, MA.

Tukey, John W. (and Mosteller, F.) (1977b). *Data Analysis and Regression: A Second Course in Statistics,* Addison-Wesley Publishing Company, Inc., Reading, MA.

Tukey, John W. (1981d) *Analiz Rezul'tatov Nablyudeniyi, (The Analysis of the Results of Observations,* Russian edition of *Exploratory Data Analysis)* translated by Ph.D. candidates A. F. Kusznira, A. L. Petrosyana and E. L. Reznikova, under the direction of B. F. Pisarenko, Mir Press, Moskova, USR, 693 pages.

Tukey, John W. (and Mosteller, F.) (1982k). *Analiz Dannykh I Regressiya,* Russian edition of *Data Analysis and Regression: A Second Course in Statistics,* translated from English by Yu, N. Blagovshchenskogo, editing and preface by Yu, P. Adleva, Financy I Statistika Press, Moskva, USSR Volume I: 319 pages, Volume II: 239 pages.

Tukey, John W. (and Hoaglin, D. C., Mosteller, F. as co-editors) (1983b). *Understanding Robust and Exploratory Data Analysis,* John Wiley & Sons, Inc., New York.

Tukey, John W. (1984b). *The collected Works of John W. Tukey, Volume I Time Series: 1949-1964,* edited by David R. Brillinger, Wadsworth, Belmont, CA.

Tukey, John W. (1985a). *The collected Works of John W. Tukey, Volume II Time Series: 1965-1984,* edited by David R. Brillinger, Wadsworth, Belmont, CA.

Tukey, John W. (and Hoaglin, D. C., Mosteller, F. as co-editors) (1985). *Exploring Data Tables, Trends, and Shapes,* John Wiley & Sons, Inc., New York.

Papers

(If the name is preceded by a * , see Supplementary References on page 11)

Bloomfield, Peter (and Anderssen, R. S.) (1974). "Numerical differentiation procedures for non-exact data," *Numer. Math.* 22.

Bloomfield, Peter (and Anderssen, R. S.) (1976). "Properties of the random search in global Optimization, *J. Optimization Theory and Applications,* 16: 383-398.

Braun, Henry (1975). "Polynomial bounds for probability generating functions," *J. Appl. Prob.* 12:507-519.

Braun, Henry (1978). "Polynomial bounds for probability generating functions II, *Z. Wahrsch. Verw. Geb.* 42: 13-21.
Also Technical Report 106.

Braun, Henry (1980a). "Testing for goodness o fit in the presence of nuisance parameters," *J. Roy. Stat. Soc. B,* 42: 53-63.

Braun, Henry (1980b). "Regression-like analysis of birth interval sequences," *Demography,* 17: 207-221.
(Technical Report 139)

Braun, Henry (with McNeil, D. R.) (1981). "Testing in Robust ANOVA" *Communications in Statistics Theory and Methods,* A10: 149-165.
Also Technical Report Nos. 94, 126, 129, 139.

Easton, George S. (with Elvezio Ronchetti) (1986). "General saddlepoint approximations with applications to L-statistics," to appear in *J. Amer. Stat. Assoc. 81.*
Also Ph.D. Thesis 1985, Technical report No. 274.

Finch, Steven J. "Robust univariate test of symmetry," JASA 72:387-392.
Also Ph.D. Thesis 1974.

Goodall, Colin (1986). "Comment" to F. L. Bookstein: 'Size and shape spaces for landmark data,' to appear *Statist. Sci.*

Gross, Alan M. (1976). "Confidence interval robustness with long-tailed symmetric distributions," *JASA,* 71: 409-416.
Also Ph.D. Thesis 1973.

Horn, Paul S. (1983). "Some easy t statistics," *JASA* 78: 930-936.
Also Ph.D. Thesis 1981.

Horn, Paul S. (1985). "On borrowing spread from auxiliary samples in the one-sample problem," *Communic. in Statist., Theory and Methods,* **14:3107-3124.**
Also Technical Report Nos. 214, 229.

Hurvich, Clifford M. (1985). "Data-driven choice of a spectrum estimate: extending the applicability of cross validation methods," *J. Amer. Statist. Assoc.* 80: 933-940.

Hurvich, Clifford M. (1986). "Data-dependent spectral window: generalizing the classical framework to include maximum entropy estimates," to appear *Technometrics,* 28:.

Kafadar, Karen (1982a). "A biweight approach to the one-sample problem," *JASA,* 77: 416-424.
Also Technical Report 151

Kafadar, Karen (1982b). "Using biweight M-estimates in the two-sample problem Part I: Symmetric Populations," Commun. Statist.. Theory. Meth., 11(17), 1883-1901.
(Technical Report 152)

Kafadar, Karen (1985). "The efficiency efficiency of the biweight as a robust estimator of location," *J. Res. Nat. Bur Stds.,* 88: 104-116.
Also Ph.D. Thesis 1979 and Technical reports Nos. 151, 152, 153, 154.

Krystinik, Katherine (Bell) (1987). The pushback, a robust location estimator, fine-tuned using configural polysampling," submitted to a technical journal.
Also Technical Reports Nos. 191, 195, 210, 211.

Morgenthaler, Stephan (1986a). Asymptotic for configural location estimators," to appear in *Ann. Statist.,* 14:.

Morgenthaler, Stephan (1986b). "Confidence intervals for a location parameter: a configural approach," to appear *J. Amer. Statist. Assoc.* 81:
Also Technical Reports Nos. 195, 252, 253, 254, 255.

Pregibon, Daryl (1980a). "Applications of resistant fitting to a class of nonlinear regression models," (Abstract) *Biometrics,* 37: 189.

Pregibon, Daryl (1980b). "Logistic regression diagnostics," *Ann Statist.* 9: 705-724.
Also Technical report No. 185.

Pregibon, Daryl (1980c). "Comments on paper by P. McCullagh," *J. Roy. Stat. Soc.* B, 42: 138-139.
Also Technical Reports Nos. 185, 186.

Ronchetti, Elvezio, (and Rousseeuw, P.) (1985a). "Change-of-variance sensitivities in regression analysis," *Zeitschrift fur W'keitstheorie und Verw. Gebiete,* 68: 503-519.

Ronchetti, Elvezio (1985b). "Robust model selection in regression, *Statistics & Probability Letters),* 3: 21-23.

Ronchetti, Elvezio (and Yen, J. H.) (1986a) "Variance-stable R-estimators," to appear in *Math. Operat. & Stat., Series Statistics.* 17:.

Ronchetti, Elvezio (and Easton, George) (1986b). "General saddlepoint approximations with applications to L-statistics," to appear in *J. Amer. Stat. Assoc.,* 81:.
(Technical Report No. 274)
Also Technical Report Nos. 266, 274.

Ronchetti, Elvezio (1986c). "Robust C (α)-type tests for linear models," to appear in *Sankhya*

Ronchetti, Elvezio M. (and Hampel, F. R., Rousseeuw, P. J. and Stahel, W. A. as co-authors) (1986d) See "Books" above.
Also Technical Reports Nos. 257, 258, 259, 266, 274.

Siegel, Andrew F. (with Mosteller, F., Trapido, E., Youtz, C) (1981). "Eye -fitting of straight lines," *The American Statistician,* 35: 150-152.

Siegel, Andrew F. (with Olshan, A. F., Swindler, D. R.) (1982a). "Robust and least-squares orthogonal mapping: Methods for the study of cephalofacial form and growth," *American Journal of Physical Anthropology,* 59: 131-137.

Siegel, Andrew F. (with Benson, R. H., Chapman, R. E.) (1982b). "On the measurement of morphology and its change," *Paleobiology,* 8(4), 328-339.

Siegel, Andrew F. (1982c). "Robust regression using repeated medians," *Biometrika,* 69: 242-244.
(Technical Report 172)

Siegel, Andrew F. (with Holst, L.) (1982d). "Covering the circle with random arcs of random sizes," *Journ. Appl. Prob.,* 19: 373-381.
cp. Technical Report 171.

Siegel, Andrew F. (1982e). "Geometric data analysis; an interactive graphics program for shape comparison," (Proceedings of the 1980 Army Research Office Workshop in Modern Data Analysis) *Modern Data Analysis,* eds. A. F. Siegel and R. L. Launer, Academic Press, New York, 103-122.

Siegel, Andrew F. (with Zambuto (later O'Brien) F.) (1983a). "Quadrature designs for unbiased estimation of the integral of any functions," *1983 Proceedings of the Statistical Computing Section of the American Statistical Association,* 69-73.

Siegel, Andrew F. (with Sugihara, G.) (1983b). "Moments of particle size distributions under sequential breakage with applications to species abundance," *J. Appl. Prob.* 20, 158-164.

Siegel, Andrew F. (1983c). "Low median and least absolute residual analysis of two-way tables," *J. Amer. Statist. Assoc. 789: 371-374.*

Siegel, Andrew F. (with Gnanadeskikan, R., Kettenring, J. R. and Tukey, P. A.) (1983d). "Exploratory data analysis," Proc. 4th Int. Conf. Math. Educat., ed. M. Zweng, et al, Birkhouser, Boston, 344-357.

Siegel, Andrew F. (and O'Brien, F.) (1984a). "Unbiased random integration methods with exactness for low order polynomials," *Proceedings of the Twenty-eight Conference on the Design of Experiments in Army Research Development and Testing*, 351-355.

Siegel, Andrew F. (with Sampson P. D.) (1984b). "The measure of 'size' independent of 'shape' for multivariate log-normal populations: Definition and applications. *Proc. XII Internat. Biometric Conference 235-244.*

Siegel, Andrew F. (1985a). "Modeling data containing exact zeroes using zero degrees of freedom," to appear in *J. Roy. Statist. Soc. B 47:*

Siegel, Andrew F. (and Sampson, P. D.) (1985b). "Consistent estimation in partially observed random walks," *J. Amer. Statist. Assoc. 85:*

Siegel, Andrew F. (and Guttorp, P.) (1985c) "Consistent estimation in partially observed random walks," *Ann. of Statist.* 13: 958-959

Siegel, Andrew F. (with Sampson, P. D.) (1985d). "The measure of 'size' independent of 'shape' for multivariate lognomial populations," *J. Amer. Statist. Assoc.*, 80: 910-914.

Siegel, Andrew F. (with Romano, J. P.) (1986). See "Books" above.

Siegel, Andrew, F., (with Pinkerton, J. R.) (1987). "Robust comparison of three-dimensional shapes with an application to protein molecule configurations," submitted to *Technometrics.*
Also Technical reports Nos. 171, 172, 173, 193, 215, 216, 217, 219, 222, 226, 237, 239.

Simon, Gary (1974). "Alternative analyses for the singly-ordered contingency table," *JASA.* 69: 971-976.
Also Technical Report No. 33.

***Note: Letters used with years on John Tukey's papers correspond to bibliographies in all volumes of his collected papers ***

Tukey, John W. (with Friedman, J. H.) (1974a). "A projection pursuit algorithm for exploratory data analysis," *C-23 IEEE Transactions on Computers*, No. 9, 881-890.

Tukey, John W. (with Beaton, A. E.) (1974c). "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *16 Technometrics*, No. 2, 147-185.

Tukey, John W. (1974g). "Named and faceless values: an initial exploration in memory of Prasanta C. Mahalanobis," *Sankhya:* The Indian Journal of Statistics, Vol 36, Series A, Pt. 2, 125-176.

Tukey, John W. (1974h). "A further analysis of the first phase of the Princeton Robustness Study: Examples of less standard two-way table analysis, *Exploring Data Analysis: The Computer Revolution in Statistics*, W. J. Dixon and W. L. Nicholson, University of California Press, Chap. 5, 229-311.

Tukey, John W. (for 1977a and 1977b) see "Books" above.

Tukey, John W. (with McCarthy, J. L.) (1978c). "Exploratory analysis of aggregate voting behavior: Presidential elections in New Hampshire, 1896-1972," *Social Science History*, 292-331.

Tukey, John W. (1978e). "The ninther, a technique for low-effort robust (resistant) location in large samples," *Contributions to Survey Sampling and Applied Statistics*, Academic Press, Inc., 251-257.

Tukey, John W. (1979c). "A study of robustness by simulation: Particularly improvement by adjustment and combination," (presented at ARO Workshop on Robustness in Statistics, April 11-12, 1978) *Robustness in Statistics*, Academic Press, New York, 75-102.

Tukey, John W. (1979d). "Robust techniques for the user," *Robustness in Statistics,'* Academic Press, New York, 103-106.

Tukey, John W. (1979f). "Comment on Emanuel Parzen's "Non-parametric statistical data modeling," *JASA*, 74: 105-131.

Tukey, John W. (1980f). Styles of data analysis, and their implications for statistical computing," *COMPSTAT 1980: Proceedings in Computational Statistics*, M. M. Barrett and D.Wishart, eds., Physica-Verlag, Vienna, 21-31. Tukey, John W. (1981d) see "books" above.

Tukey, John W. (with Tukey, P. A.) (1981e). "Graphical display of data sets in 3 or more dimensions," Chapters 10, 11 and 12 *Interpreting Multivariate Data*, ed. V. Barnett, Chichester: Wiley, 189-275.

Tukey, John W. (with Mosteller, F.) (1982a). "Combination of results of stated precision: I. The optimistic case," *Utilitas Mathematica*, 21, May/June, Winnepeg, Canada, 155-178.

Tukey, John W. (1982b). "Another look at the future," *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, eds. Karl W. Heiner, Richard S. Sacher and John W. Wilkinson, Springer-Verlag, New York, 2-8.

Tukey, John W. (with Tukey, P. A.) (1982c). "Some graphics for studying 4-dimensional data," *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, eds. Karl W. Heiner, Richard S. Sacher and John W. Wilkinson, Springer-Verlag, New York, 60-66.

Tukey, John W. (with Mosteller, F.) (1982e). "Combination of results of stated precision: II. A more realistic case," *Papers in Honor of W. Cochran*, eds. P.S.R.S. Rao and Joseph Sedransk), Wiley, New York, 223-252.

Tukey, John W. (1982f). "Discussion," (The Role of Statistical Graduate Training), *Teaching of Statistics and Statistical Consulting*, eds. J. S. Rustagi and D. A Wolfe, Academic Press, Inc., 379-389.

Tukey, John W. (with Seheult, A.) (1982i). "Some resistant procedures for analyzing 2^n factorial experiments," *Utilitas Mathematica, 21B*, May, Winnepeg, Canada, 57-97.

Tukey, John W. (with Mallows, C.) (1982j). "An overview of techniques of data analysis, emphasizing its exploratory aspects," *Some Recent Advances in Statistics*, eds. J. Tiago de Oliveira, et al, London: Academic Pres, 111-172.

Tukey, John W. (1982k). See "Books" above.

Tukey, John W. (with Mallows, C. L.) (1982l). "An overview of techniques of data analysis, emphasizing its exploratory aspects," *Some Recent Advances in Statistics*, eds. J. Tiago de Oliveira, and Benjamin Epstein, Lisboa, Publicacoes do II Centenario da Academia das Ciencias de Lisboa, (limited national edition) 111-172. (Another edition of 1982j).

Tukey, John W. (1982m). "Introduction to styles of data analysis techniques," Proceedings of the 1980 ARO Workshop on Modern Data Analysis, *Modern Data Analysis; Proceedings* eds. R. Launer and A. F. Siegel, Academic Press, New York 1-11.

Tukey, John W. (1982n). "The use of smelting in guiding re-expression," Proceedings of the 1980 ARO Workshop on Modern Data Analysis, *Modern Data Analysis; Proceedings*, eds. R. Launer and A. F. Siegel, Academic Press, New York, 83-102.

Tukey, John W., (1983a). "The relationship of empirical analysis to more narrowly modeled analysis," Appendix A in *Age, Time and Fertility: Applications of Exploratory Data Analysis*, Mary B. Breckenridge, Academic Press, 174-281.

Tukey, John W. (1983b). See under "Books" above.

Tukey, John W. (1984a). "Styles of spectrum analysis," *A Celebration in Geophysics and Oceanography—1982, in Honor of Walter Munk*, Scripps Institute of Oceanography, Reference Series 84-5, La Jolla, CA, 100-113. [Also in the *Collected Works of John W. Tukey, Volume II: Time Series, 1965-1984*, Wadsworth Publishing Company, Belmont, CA 1985, 1143-1153.]

Tukey, John W. (1984b). See under "Books" above.

Tukey, John W. (1985a). See under "Books" above.

Tukey, John W. (1985b). See under "Books" above.

Tukey, John W. (with Tukey, P. A.) (1985k). "Computer graphics and exploratory data analysis: An introduction," (Proceedings of the Sixth Annul Conference and Exposition, held at Dallas Convention Center, Dallas, Texas, April 14-18, 1985), *Computer Graphics "85, Conference Proceedings, Vol. III*, 773-785.

Tukey, John W. (1985l). "Improving crucial randomized experiments - - especially in weather modification - - by double randomization and rank combination, *Proceedings of the Berkeley Conference in Honor of Jerzy* Neyman and Jack Kiefer, eds. Lucien M. Le Cam and Richard A. Olshan, Wadsworth Publishing Company, Belmont, CA., 330-359.

Tukey, John W. (1985m). "Discussion of Peter Huber's projection pursuit," *Ann. Stat.*, 13: 517-518.

Tukey, John W. (with Hoaglin, D. C) (1985n). "Checking the discrete distributions," *Exploring Data Tables, Trends, and Shapes*, John Wiley & Sons, Inc. Publishers, New York, Chapter 9: 345-416.

Tukey, John W. (1985o). "The variance of slopes of lines fitted to groups: an analysis of the Johnstone and Velleman Monte Carlo results," *J. Amer. Statist. Assoc.*, 80: 1055-1059.

Tukey, John W. (with Iglewicz, B. and Hoaglin, D. C.) (1986**). "Performance of some resistant rules for outlier labeling," submitted to *JASA*.

Tukey, John W. (1986a). "Sunset salvo," (based on material presented or handed out at the 1985 Spring Symposium of the Northern New Jersey Chapter of ASA, May 6, 1985) *The American Statistician*, 72-76.

Tukey, John W. (1986*). "Configural polysampling," (presented as the John von Neumann Lecture for 1985 at the SIAM national meeting in Pittsburgh, June 24-26, 1985), to appear in *SIAM Review*.

Tukey, John W. (with Johnson, Eugene G.) (1986**). "Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao data," (to appear in a book in honor of Cuthbert Daniel's 80th birthday), Wiley & Sons, Inc. Publishers, New York.

Tukey, John W. (1986****). "Limited randomization with detailed reassignment as the key to taking advantage of modern summaries," to appear in *Proceedings of the 30th Conference on the Design of Experiments in Army Research, Development and Testing*, Army Research Office, Durham, N. C.
Also Technical Reports Nos. 129, 143, 252, 272, 275, 288.

*** Unpublished papers (1974-86, ARO) by John W. Tukey presently planned for publication in future volumes of his collected works ***

Tukey, John W. (1974U5) "Notes on the swindles for locations and scale"

Tukey, John W. (1974U7) "On widthing, a preliminary list of some estimators"

Tukey John W. (1976U3) "Notes on the mean square successive differences as squared denominators"

Tukey, John W. (1979U1) "Introduction to guided re-expression"

Tukey, John W. (1980U6) Suggestions for an analysis of intermediate formality for interlap trials, (attachment to long letter to Dr. Marian Scott, U. of Glasgow, May)

Tukey, John W. (1980U12) "Steps toward a universal univariate distribution analyzer"

Tukey, Jonn W. (1981U1) "A more rational approach to analyzing our 5x5x5 and related matters, emphasizing symmetry"

Tukey, John W. (1981U5) "Choosing techniques for the analysis of data"

Tukey, John W. (1981U6) "Do derivations come from heaven"

Tukey, John W. (1982U2) "More on comparison of quadrature formulas"

Tukey, John W. (2014) "Indirect measurement and blunder resistance"

* * *

Tyler, David E. "Asymptotic inference for vectors," *Ann. Statist.*, 9: 725-736.
Also Ph.D. Thesis 1979.

Velleman, Paul (1977). Robust nonlinear data smoother: definitions and recommendations," *Proc. Nat. Acad. Sci.*, 74: 434-436.

Velleman, Paul (1980). "Definition and comparison of robust nonlinear data smoothing algorithms," *J. Amer. Statist. Assoc.*, 75: 609-615.

Velleman, Paul (with I. Johnstone) (1985). "The resistant line and related regression methods," *J.Amer. Statist. Assoc.*, 80: 1041-1054.
Also Ph.D. Thesis 1975.

Supplementary References
(not issued in 1974—86 or not under ARO sponsorship)

*Andrews, D. F. Bickel, P. J. Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, John W. (1972e). *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton, 373 pp.

*Larntz, Kinley 1978. "Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *J. Amer. Statist. Assoc.*, 73: 253-263.

*Tukey, John W. (with M. F. Freeman) (1950h. "Transformations related to the angular and the square root," *21 Ann. Math. Statist.* 607-611.

*Tukey, John W. (1971a). *Exploratory Data Analysis*, Volume III, limited preliminary edition, Addison-Wesley, Reading, MA.

*Tukey, John W. (with Brillinger, D. R. and Jones, L W.) (1978g). "The role of statistics in weather resources management," *11 The Management of Weather Resources*, Washington Government Printing Office.

3. Theses

Ph.D. Theses

1974—

Steven Finch, "Univariate robust test of symmetry"

Lincoln Polissor, "Parameterizing age distributions of death by cause"

1975—

Paul Velleman, "Robust non linear data smoothers - theory, definitions, and applications"

Karen Kafadar, "Robust confidence intervals for the one- and two-sample problems

Michael D. Schwarzschild, "New observation-outlier-resistant methods of spectrum estimation

David E. Tyler, "Associated asymptotic distribution theory"

1981—

Katherine Bell Krystinik, "Data modifications based on order; pushback; a configural polysampling approach"

Paul S. Horn, "On simple robust confidence procedures"

1983—

Stephan Morgenthaler, "Robust confidence intervals for location and scale parameters: The configural approach"

1984—

Dhammika J. Amaratunga, "Pushing back regression coefficients and evaluating performance via orthogonal samples"
O'Brien, Fanny L., "Polyefficient and polyeffective simple linear regression estimators and the absolute polyefficiency of the biweight regression estimator"

1985—

Clifford M. Hurvich, "A unified approach to spectrum estimation: objective estimate choice and generalized spectral windows"

George S. Easton, "Finite-sample and asymptotic approaches to compromise estimation including compromise maximum-likelihood estimators"

Senior Theses

Andrew Bruce
David A. Lax 1975
Cindy Stoughton 1985

4. Technical Reports 1974--1986

Technical Reports

Number	Title	Author and date
33	Alternative analyses for the singly-ordered contingency table	G. Simon April 1973 Revised January 1974
92	The widthing Monte Carlo Program: A computer program for a Monte Carlo Study on robust estimators on width design considerations, instructions on use, and instruction on modification	David A. Lax David J. Pasta September 1975
93	An interim report of a Monte Carlo study of robust estimators of width	David A. Lax August 1975
94	Robustness and the jackknife	Henry Braun August 1975
106	Polynomial bounds for probability generating functions, II	Henry Braun February 1976
126	A comparative study of models for reliability growth	Henry Braun J. M. Paine July 1977
129	Further progress on robust/resistant widthers	Henry Braun Michael Schwarzschild John W. Tukey
139	Regression-like analysis of birth interval sequences	Henry Braun July 1979
143	An empirical higher-rank analysis model of the age distribution of fertility	M. Breckenridge (John W. Tukey) May 1978
147	Fortran subroutines for M estimators in the linear model	A. Marazzi April 1979
151	A biweight approach to the one-sample problem	Karen Kafadar August 1979
152	Using Biweights in the two-sample problem	Karen Kafadar August 1979
153	Formulas for a two-sample Monte Carlo swindle	Karen Kafadar August 1979
154	A robust confidence interval for	Karen Kafadar

	samples of five observations	August 1979
171	Covering the circle with random arcs of random sizes	Andrew F. Siegel L. Holst September 1980
172	Robust regression using repeated medians	Andrew F. Siegel September 1980
173	Teaching robust methods for exploratory data analysis	Andrew F. Siegel October 1980
183	Eye fitting of straight lines	Andrew F. Siegel F. Mosteller E. Trapido C. Youtz January 1981
185	Assessing the behavior of robust estimates of location in small samples: Introduction to configural polysampling	Daryl Pregibon John W. Tukey March 1981
186	The second representing function for compound situations	Andrew Bruce Daryl Pregibon John W. Tukey March 1981
189	Some advanced thoughts on the data analysis involved in configural polysampling directed toward high performance estimates	John W. Tukey April 1981
190	Kinds of polyconfidence limits for centers and some thoughts on identification and selection of confidence procedures using configural polysampling	John W. Tukey April 1981
191	Some computational details of configural sampling methods	Katherine Bell (later Krystinik) Daryl Pregibon April 1981
193	Geometric data analysis: an interactive graphics program for shape comparison	Andrew F. Siegel Arpil 1981
195	Comparison of the bioptimal curve with curves for two robust estimates	Katherine Bell (later Krystinik) Stephan Morgenthaler Corrected October 1981
210	Pre-configural polysampling pushback performance	Katherine Bell (later Krystinik)

September 1981

211	Post-configural polysampling pushback performance	Katherine Bell (later Krystinik) September 1981
214	Some easy "t"-statistics	Paul S. Horn November 1981
215	Moments of particle size distributions under sequential breakage with applications to species abundance	Andrew F. Siegel G. Sugihara November 1981
216	Platonic plate tectonics: of the distribution of "triple points" on the earth's surface	A. J. Arnold Andrew F. Siegel November 1981 ce
217	Robust comparison of three-dimensional shapes with an application to protein molecule configurations	J. R. Pinkerton Andrew F. Siegel March 1982
222	Robust and least squares orthogonal mapping methods for the study of cephalogacial form and growth	D. R. Bwindler A. F. Olshan Andrew F. Siegel February 1982
224	On the measurement of morphology and its changes	R. H. Benson R. E. Chapman Andrew F. Siegel March 1982
226	Quadrature designs for unbiased estimation of the inegral of any function that are exact for law order polynomials	Andrew F. Siegel Fanny Zambuto, (later O'Brien) May 1982
229	On schemes for borrowing denominators	Paul S. Horn June 1982
237	Rarefaction curves	Andrew F. Siegel December 1982
239	Modelling data containing exact zeros using zero degrees of freedom	Andrew F. Siegel March 1983
252	The use of doubling sampling in studying robustness	Stephan Morgenthaler John W. Tukey September 1983
253	Strong confidence intervals: a compromise between the Gaussian and the Slash	Stephan Morgenthaler November 1983
254	Bi- and poly-optimized confidence	Stephan Morgenthaler

	limits for a location parameter	November 1983
255	Robust confidence intervals for a scale parameter: a compromise between the Gaussian and the Slash	Stephan Morgenthaler September 1983
257	Bounded influence inference in regression	Elvezio Ronchetti February 1984
258	Robust $C(\alpha)$ -type tests for linear models	Elvezio Ronchetti February 1984
259	Robust model selection in regression	Elvezio Ronchetti February 1984
266	Variable stable R-estimators	Elvezio Ronchetti J. H. Yen May 1984
272	The variance of slopes in line fitted to groups: An analysis of the Johnstone and Velleman Monte Carlo	John W. Tukey August 1984
274	General saddlepoint approximations with applications to L-statistics	George S. Easton Elvezio Ronchetti September 1984
275	Checking the shape of discrete distributions	D. C. Hoaglin John W. Tukey August 1984
277	Optimal Pitman regression-and-scale invariant estimators for individual p-wild Gaussian situations	Fanny O'Brien September 1984
278	Optimal Pitman variances for individual *p-wild Gaussian sampling situations* - - sample sizes 10 and 20	Fanny O'Brien September 1984
279	Compromising over p-wild Gaussian situations in a simple way: Polyeffective regression estimators	Fanny O'Brien September 1984
282	The absolute polyefficiency of the biweight regression estimator	Fanny O'Brien September 1984
288	*Sunset salvo*	John W. Tukey May 1985

5. Broadbrush account of work: 1974 — 1986

The work of this project has been mainly focused on "procedures" - - on things to do with data - - on data-analysis techniques. The main heads under which we shall mention, in broad outline, what has been done, are five:

a) exploratory and graphical procedures

b) robust/resistant procedures

c) regression procedures

d) analysis-of-variance procedures, including factorialization and multiple comparisons

e) spectrum analysis procedures

f) randomization in experimentation.

(These heads include a large fraction of the most important procedures that are applied to data. In the cases of (a), (b), (d) and (e), at least, where we stood at the close of 1973 had been significantly influenced by work earlier at Princeton involving the principal investigator.) The subsections that then follow mention a diversity of topics.

* general techniques and philosophy *

With a broad interest as have just been sketched, it would have been surprising if the work of the project had not included some general accounts, crossing over the subdivisions just noted. We sketch them here.

In 1979-80, C. L. Mallows and John Tukey prepared a rather general account of data analysis, including several novel ideas and concepts (Tukey (with CLM) 1982j); also published as Tukey (with CLM) 1982l). In 1980, John Tukey presented an account of styles of data analysis (Tukey 1982m). In 1981, he discussed the future as the keynote speaker at the 14th meeting of the "Interface" (Tukey 1982b).

In 1983, he discussed the relation between empirical (i.e. exploratory) analysis and narrowly modelled analysis (Tukey 1983a).

In 1985, he discussed a variety of general issues in the form of a "Sunset salvo" (Tech Rep. 288, Tukey 1986a)

5a. Graphical and exploratory procedures

The connection between graphical and exploratory procedures is, and will remain, close. Besides their try-it-and-see attitude and their desire to see what might be so – rather than being limited to what can be shown beyond a more or less reasonable doubt - - exploratory procedures are intended to help us see things that we did not expect to be there. In this task, graphical presentation has been our main stay. Tables of preplanned numbers, or of formulas, are not good tools to reveal the unanticipated. Pictures can do just that.

Equally, graphic displays encourage exploration.

In the relatively near future, we can expect the rapidly falling costs of computation and the equally rapid expansion in diversity of the ways in which they can look at data to make computer-scanning an important input alongside human scanning of pictures. But the computer's report of what it has dredged up is almost certain to be in the form of a picture.

* structure *

We shall mention work in this area under six heads:

- graphical techniques *per se*
- the pre-book and book phases of EDA
- pushback
- other exploratory techniques
- cognostics.
- interactive analysis

* graphical techniques per se *

(This work appears only in the later part of the 12 years covered by this report, since earlier work at Princeton had a different sponsor.

In 1979-80, P. A. Tukey and John Tukey prepared, gave and published a 3-lecture series on graphical methods for interpreting data in 3 or more dimensions (Tukey (with PAT) 1981e). These chapters incorporated a number of novel approaches.

In 1981-82, they prepared, presented and published an account focused on 4-dimensional data (Tukey (with PAT) 1982c).

In 1983-85, D. C. Hoaglin and John Tukey prepared and published an account of graphical techniques for comparing observed sequences of counts with the standard discrete distributions, proposing a number of new approaches (Tukey (with DCH) 1983d).

During 1984 and 1985 John Tukey and P. A. Tukey developed three sorts of "frames" (using wooden strips, transparencies, and linking devices) to allow dynamic views to be projected with an ordinary overhead projector. Both layered skewing and alternation can be shown, each by an appropriate device (unpublished, but publicly demonstrated).

In 1985 they prepared, presented and published an introduction to "Computer Graphics and Exploratory Data Analysis" (Tukey (with PAT) 1985k).

In 1985, Eugene Johnson and John Tukey prepared an account of an exploratory, graphical approach to factorial (i. e. 2 or more coordinates in all combinations) data - - to the sort of data usually treated in terms of the classical analysis of variance (Tukey (with EJ) 1986***).

* Exploratory Data Analysis *

During 1974-76, much effort was applied to new techniques of exploratory data analysis. These were mainly reported in the First Edition of John Tukey's *Exploratory Data Analysis*. (Tukey 1977a).

Applications to demography were reported by Mary Breckenridge (Tech. Rep. 143) and, later, a book, (Breckenridge 1983), with an appendix by John Tukey, (Tukey 1983a).

Applications to semi-markov process data were studied by Henry Braun, leading to regression-like analysis of birth intervals (Braun 1980b).

Applications to voting behavior were studied by J. L. McCarthy and John Tukey (Tukey (with JLM) 1978c).

* pushback *

The use of "pushback" in which individual observed values are modified in the light of the presence of other values has been considered, at intervals throughout the twelve years 1974—86.

The results of early work by John Tukey on the question of "deblurring" an observed distribution - - doing one's best to eliminate the effects of, say, measurement error were published in 1974 (Tukey 1974g).

Work, in 1974-75, by L. F. Nanni and John Tukey focussed on two uses - - as a contribution to robust centering and as a route to a plot that would do better what a conventional probability plot would do (unpublished).

In 1980-81, Katherine (Bell) Krystinik carried through a study of the first use, finding that simple summaries of pushed-back values showed high performance, and preparing a Ph.D. Thesis.

In 1986, John Tukey returned to this topic, and suggested a variety of promising modifications (unpublished).

* other exploratory techniques *

Over a period of time, D. C. Hoaglin, B. Iglewicz and John Tukey have studied to quantitative null behavior, for both Gaussian and long-tail parents, of the "fence-and-outside" labelling procedure proposed in *Exploratory Data Analysis. (Tukey (with DCH, BI) 1981a and 1986**).*

In 1985-6, Katherine Hansen and John Tukey studied more sophisticated approaches to clustering - - to dividing a given set of points into "clusters" on the basis of their mutual distances. By avoiding barriers of required simplicity, it has been possible to greatly improve the sensitivity of such a procedure. Present techniques separate moderately overlapping Gaussian distributions almost as well as the linear discriminants would that could be calculated only if the true distributions were known. Work continues (unpublished, some aspects reported in a talk to the North American Classification Society).

* cognostics *

Computer *interpreted* diagnostics seem almost certain to be the handmaiden and supplement to graphical display.

Attention to cognostics was first given in 1981 (Tukey 1981U_). A set of plausible suggestions were made in 1985 (Tukey (with PAT) 1985k).

* interactive analysis *

Motivated by the need to display effectively what has already been tried on a data set, John Tukey has prepared several drafts of an account of what operations might be provided and how the history of their use might be displayed (work in progress).

5b. Robust/resistant procedure

At the close of 1973, the Princeton Robustness Study (*Tukey, (with DFA, PJB, FRH, PJH, WHR) 1972c) had circulated for over a year. Very considerable progress had been made on the simplest problem: estimating a "center" from a single batch of numbers in a robust - - and consequently, would seem, in a resistant way. Most of the improvements found in the study can be traced to the results of empirical trials, either directly or through careful thought applied to understanding these results. The twelve years that have followed at Princeton have seen (i) an extension of the procedures to a variety of problems, (ii) the development of configural polysampling which enables us to bound the possible (in finite-sized samples) and (iii) recent asymptotic developments, which promise a new era of broadening of applications.

At the close of 1973, the early beginnings of robust (non-linear) smoothing were in place. (The start may have been in the Limited Preliminary Edition of *Exploratory Data Analysis* (*Tukey 1971a).)

We shall review project activity here under 8 heads:

- robust (non-linear) smoothing (1974-1986)
- extended problems (1975-80)
- empirically-based improvement techniques (1977-76)
- uses of order statistics (1979-83)
- robust shape comparisons (1980-83)
- configural polysampling (1980-86)
- new problem types (1982-86)
- asymptotic improvement techniques (1983-86)

In addition to what is reported here in 5b, large parts of what is reported below under 5c (regression), 5d (analysis of variance), and 5e (spectrum analysis) involve the use of robust-resistant concepts so deeply is to fit in here in 5b had we wished. And the importance of the procedures in 5f stems from the possibility of using robust/resistant summaries. We have chosen the actual structure, however, in order to keep the story as simple as we know how to do.

The relation to 5a is also closer than one might think. While the emphases are quite different - - "finding appearances" in 5a, "being stringent in diverse circumstances" here - - good examples of exploratory procedure have to have a good dose of robustness, even though extreme high stringency usually need not be sought. (As we move to deeper involvement with modern computers, greater stringency that involves no other cost than computation is more and more likely to be seized upon.)

This latter convergence is to be seen rather clearly in the combined, and on occasion even interlaced, treatment of exploratory and robust techniques in the two recent books - - *Understanding Robust and Exploratory Data Analysis* and *Examining Data Tables, Trends, and Shapes* - - edited by D. C. Hoaglin, F. Mosteller, and John Tukey (Tukey (with DCH and FM) 1983b, 1985b).

* robust (non-linear) smoothing *

Smoothing by moving linear combinations was the classical form of smoothing, suffering from (i) too much attention to outliers and (ii) filling in valleys and cutting down hills. Use of non-linear moving combinations can ameliorate both these difficulties. It still does "smoothing by value-change," giving us a smoothed value in place of each initial value. In certain applications, there is use for a different process - - "smoothing by excision" - - in which we set aside some of the initially given values.

The use of robust smoothing as a way to robustify the fitting of polynomials was developed and illustrated by A. E. Beaton and John Tukey (Tukey with AEB) 1974c). Improved techniques of robust smoothing were developed, and reported in the first edition of *Exploratory Data Analysis* (Tukey 1977a).

The role of "head banging" as a fundamental concept in the construction of robust/resistant smoothers was recognized in 1978 and extended to smoothing in the plane (cp. Tukey (with PAT) 1981e).

Re-expression of the numerical responses (or numerical circumstances) we are analyzing - - as, in the simplest case, by taking logarithms - - is often important. We would usually like to have the choice made robustly. In many situations it is natural to "guide" the re-expression by a rank related version of the numbers at hand. Typically we then want to be guided in the large-scale behavior of the result, but to reflect the small-scale behavior of the original numbers. A procedure for doing this, emphasizing smoothing by excision and based on starting to smooth divided differences (that connect initial values to guiding values), is called *smelting* and leads to pictures which can guide the choice of simple-formula re-expressions. (Tukey 1982n).

Recently, John Tukey has been reviewing and extending the available robust smoothing techniques (Technical Report in preparation).

* extended problems *

In the latter half of the 70's, the procedures, and the background leading to their choice, for the one-sample problem of center finding were extended to other related problems.

The problem of robust/resistant assessment of *width of distribution* was addressed in 1975-77 (Tech. Rep. 129 (Tukey (with FM) 1977b). Further small improvements were made later (unreported).

Work directed toward improved empirical assessment of stability for widthers was begun in 1977 and then interrupted.

The extensions of one-sample robust-resistant estimates to *two-sample comparisons* (differences) of centers was successfully undertaken in 1978-79 by Karen Kafadar (Ph.D. Thesis 1979, Tech. Reps. 152, 154 Kafadar 1982a, 1982b, 1985). She also showed that both the interval estimates proposed by Gross in the one-sample case, and her intervals for differences still behaved very well for confidence coefficients *very close to unity (tail areas like 0.01%).*

An approach to robust *correlation* based on robust widthing was suggested in the book by F. Mosteller and John Tukey (Tukey (with FM) 1977b).

The bias of repeated-median correlation estimation was assessed (by simulation) by Andrew Siegel in 1980 (unpublished).

For extensions to *regression* see 5c and for extensions to *analysis of variance* see 5d.

* empirically-based improvement techniques

Methods of improving the performance of robust estimators by combining two or more, using each in its prescribed part of the configuration space, were developed and reported

(Tukey 1979c).

* use of order statistics *

The *ninther*, technique for rapid center estimation "on the fly" when dealing with very large samples was developed and reported (Tukey 1978e).

The behavior of order statistics from the 3 original corner distributions was studied by Andrew Bruce (senior thesis, supervised by Daryl Pregibon) who later extended his results to 2 additional stretched-tail distributions. Among linear combinations of order-statistics, his results favor the use of the so-called ab-mean. (unpublished). Related material, involving the "2nd representing function", was reported in 1981 by Andrew Bruce, Daryl Pregibon and John Tukey (Tech. Rep. 186).

The possibilities of "easy-t" confidence intervals based upon only 2 order statistics, actual or mid-interpolation, for sample sizes of 5 to 20 were studied and reported by Paul Horn (Ph.D. thesis 1981; Horn 1983, Horn 1985; Tech. Rep. 229).

* robust shape comparison *

The comparison of shapes of geometric figures (including projections of shapes of animals) is often better conducted when the rescalings and rotations implied by "shapes" are done robustly. Work on such methods, using repeated medians, had been initiated by Andrew Siegel before coming to Princeton. Extensions to the three-dimensional case were studied by John Pinkerton (Junior Paper under Siegel's direction, Tech. Rep. 217, Siegel and Pinkerton 1986).

Both repeated-median and least square methods were computerized by Siegel (Tech. Rep. 193, Siegel 1982e). The usefulness of such techniques was demonstrated on an anthropological example by A. F. Olshan, Andrew Siegel, and D. R. Swindler (Tech. Rep. 222, Siegel (with AFO, 1982a).

Shape and pattern matching were reviewed by R. H. Benson, R. E. Chapman and Andrew F. Siegel (Tech Rep. 224, Siegel (with RHB and REC) 1982b).

The unique decomposition of a multivariate log-normal population into statistically independent shape and size variates was found by P. D. Sampson and Andrew Siegel (Siegel (with PDS) 1984b, 1985d).

* growth modelling *

Colin Goodall has developed (i) a growth model comprising a continuously-varying deformation tensor field with explicit components for measurement error, shape change, and inter-individual variation and (ii) a finite-element model involving cell-reinforcement polarity, cell-division direction, and cell deformation. (Work in progress).

* configural polysampling *

The most clearly defined estimation problems are those in which an invariance requirement ensures that no personal prejudice or external information about which estimate values are likely is involved in the results. (External information about other matters, such as shapes of distribution likely or possible can, and often should be included.) The developments of the Princeton Robustness Study (*Tukey (with DFA, PJB, FRH, PJH, WHR) 1972e) left the finite-sample - - realistic - - study of robust estimation as something focused on selected examples of distribution shape, often as few as two.

In the centering (location) problem, a configuration consists of all sets of values $\{y_i\}$ which differ from one another only by location and scale - - that is all of the form $\{A + by_i\}$ for fixed $\{y_i\}$ and any a and any $b \neq 0$ (or, sometimes, any $b > 0$). Here the natural invariance requirement is an equivariance one, namely

estimate from $\{a + by_i\} = a + b \cdot$estimate from $\{y_i\}$

Prior to the introduction of configural polysampling, which began in 1980, attempts to optimize centering performance for a particular sample size and, say, two particular distribution shapes, involved

- many repetitions of inventing plausible estimates
- drawing separate sets of samples from each distribution shape,
- evaluating the performance of each estimate at each sample (usually for a family of related samples, but usually not for a whole configuration), still for each situation separately,
- plotting the results, with performance for each shape labelling the corresponding axis,
- looking at the plot, to guess both where the boundary between the possible and the impossible lay and what kind of a further modified estimate would bring us closer to the boundary.

The Princeton Robustness Study and its follow on involved this sort of trial for 700-odd estimates (and several thousand 50-50 mixtures of estimate pairs). We still had no clear idea where the boundary fell.

The introduction of configural polysampling, which began in 1980, made it possible to work with a single set of configurations, applicable when appropriately weighted to each of the shapes concerned. (Polysampling refers to a single set of samples which, when used with different weights, provide weighted random samples from two or more populations.) Here, working with a configuration for a shape means evaluating, by numerical integration, a small number of two-dimensional integrals. The combination of (i) these numerical integration results and (ii) "shadow prices" for the shapes, allows us to determine the values of optimized estimates at each configuration used, and the corresponding conditional variances, one per shape. The weighted sampling of configurations extends this result - - subject to sampling error, as in all forms of simulation - - to the unconditional variances, still one per shape. Each set of shadow prices lets us estimate a point on the boundary between the possible and the impossible. Taking the shadow prices as parameters, we can estimate the whole run of the boundary.

Work on the fundamental ideas and formulas began in 1980 and was reported by Daryl

Pregibon and John Tukey in 1981 (Tech. Rep. 185). Extensions and related matters were reported by John Tukey (Tech. Rep. 189) and by Katherine Bell (later Krystinik) and Daryl Pregibon (Tech. Rep. 191). An illuminating comparison of two standard M-estimate (each both fully iterated and one-step) was made by Katherine Bell (later Krystinik) and Stephan Morgenthaler (Tech. Rep. 195).

Michael Cohen worked on improvement of simple estimates by regression adjustments guided by optimum values obtained by configural polysampling in 1981-82 (unreported). John Tukey worked on other improvement schemes in 1982 (unreported).

Stephan Morgenthaler initiated work on optimization for two end shapes and an intermediate shape in 1981. This work was continued by Michael Cohen, George Grover, Stephan Morgenthaler and John Tukey (not reported in detail, but see Tukey 1986*).

Applications of double sampling to improve computational efficiency were studied by Stephan Morgenthaler and John Tukey (Tech. Rep. 252).

For applications of configural polysampling to regression see section 5C below.

A review paper on configural polysampling was presented to the Society for Industrial and Applied Mathematics as the 25th von Neuuman lecture, and will appear in *SIAM Review* (Tukey 1986*).

In order to make available an adequate account of configural polysampling, Stephan Morgenthaler and John Tukey have undertaken the editing of a volume on this subject, with chapters contributed by all of those named above.

* new problem types *

J. Pederson, with Daryl Pregibon's guidance, wrote a junior paper (1980) on small-sample properties of resistant estimates in non-location situations.

Possibilities for alternative definitions of *confidence limits* in a configuration-oriented robust world were suggested by John Tukey (Tech. Rep. 190).

The actual situation for confidence intervals was investigated in some detail by Stephan Morgenthaler (Ph.D. Thesis 1983, Tech. Reps. 253, 254, 255, Morgenthaler 1986b) who examined confidence intervals for width (for scale) as well as for center (for location).

Robust estimation in non-linear models is being studied by Elvezio Ronchetti and S. Morgenthaler.

Robust signal detection has been studied by Elvezio Ronchetti and M. Weiss (1983-4, unreported).

Robust model selection is discussed under 5c, regression procedures.

Asymptotics for configural location estimates have been discussed by Stephan Morgenthaler (Morgenthaler 1985a) and George Easton (Ph.D. Thesis 1985).

* advanced improvement techniques *

Optimizing asymptotic variance of an R-estimator while bounding its sensitivities to gross errors and changes of variance was studied by Elvezio Ronchetti and J. Yen (Tech. Rep. 266, Ronchetti and Yen 1986a).

The relation of *small-sample asymptotics* to bootstrapping has been studied by Elvezio Ronchetti (1983-85, work in progress).

The use of small-sample asymptotics to find approximations to the *density of trimmed means* has been explored by Elvezio Ronchetti and George Easton who have extended the applicability of saddle-point approximations to general statistics, including linear combinations of order statistics (Tech. Rep. 274, Easton and Ronchetti 1986).

George Easton has studied the asymptotic behavior of the optimal compromise estimates for location. The approximations required led to a new class of compromise estimators (CMLE's) that are based on the likelihood functions of the chosen situations. Finding these (closely approximately optimum) estimates is much simpler - - a biparameter optimization instead of a two-dimensional numerical integration - - than finding the exactly optimum ones. Their performance is very promising; they should be extendible to problems we do now know otherwise how to attack. (Ph.D. Thesis, 1985).

George Eastor has also found that a simple modification to the likelihood function of the slash distribution substantially improves the performance of both its (pseudo) MLE and (pseduo) CMLE's (Ph.D. Thesis, 1985).

* expositions and weeks *

Robust methods were expounded for the user John Tukey (Tukey 1979).

The teaching of robust methods was expounded by Andrew Siegel (Tech. Rep. 173, Siegel (with (RG, JK, and PAT) 1983c).

Robustness weeks, where selected active workers in the field could interact effectively, were held in Princeton 15-19 March 1980 and 4-8 May 1981.

5c. Regression procedures

Work here is moderately diverse, and is discussed under these heads:

- functional issues
- repeated medians
- regression in general classes
- fitting straight lines
- strengthening theoretical understanding.

A general account of some elementary, but vital, issues in regression are given in *Data Analysis and Regression* (Tukey (with FM) 1977b). A variety of topics are discussed in *Exploring Data Tables, Trends and Shapes* (Tukey (with DCH, FM) 1985b).

* functional issues *

A formalized fitting (regression) problem has two parts:

- the functional model, which describes the alternatives from which the fit is to be chosen, and
- the stochastic model, which describes how the error might be distributed - - best in terms of several alternatives, each ordinarily involving at least on variability parameter.

It will clearly be increasingly important to bring into the functional model flexibility analogous to what robust/resistant techniques have put into the stochastic model.

One way to do this is to consider what are the natural generalizations of, say, $a + bx$. The answer, as John Tukey had recognized by 1979, was usually not to go to polynomials of degree 2 or more. Rather, in most situations, other simple functions preserving monotonicity rather than preserving additive appearance of the constants, are a natural choice.

If we consider, for flexibility, two (or more) alternative families of possible fits, neither including the other, it is not usually sensible to estimate a single parameter (or a single set of parameters) for both. But it may be not only possible but desirable to ask if fitting either (possibly both) shows a significant improvement in the remaining residuals. If the families are rather similar, the usual (or unusual) tests for fitting either separately will be highly correlated. Thus a Bonferroni calculation will be grossly overconservative. If we are dealing with an experimental situation where empirical randomization (perhaps within a limited randomization) can be used, we can easily assess the significance of any convenient combination of the two or more tests (Tukey 1985l).

More attention to such functional problems is needed.

* repeated median *

Beginning in 1980, Andrew Siegel studied and reported on a technique of repeated median fitting (Tech. Rep. 172, Siegel 1982c). This technique is quite useful for 2-parameter problems, like fitting straight lines, but rapidly demands heavy computations as the number of parameters increases. The relationship of repeated median fitting to other types of fits has been discussed by I. M. Johnstone and Paul Velleman (Velleman (with IMJ) 1985).

* regression in general classes *

In 1980, Daryl Pregibon prepared and gave, at Perugia, a five-lecture series on data-analytic methods for fitting generalized linear models (unpublished).

In 1980, Daryl Pregibon reported on techniques for dealing with logistic regression problems (Pregibon 1980a, 1980b) and discussed McCullagh's work on related topics (Pregibon 1980c).

In 1983, John Tukey and Paul Velleman prepared a draft account of what a reasonable computer-supported regression procedure might include (unpublished).

Robust model selection in regression has been studied by Elvezio Ronchetti (Tech. Rep. 259, Ronchetti 1985).

A plausible, highly heuristic proposal on how to start a robust-regression was made by John Tukey in 1984 (unpublished).

* fitting straight lines *

The eye-fitting of straight lines was studied by F. Mosteller, Andrew Siegel, E. Trapido, and C. Youtz (Tech. Rep. 183, Siegel (with FM, ET, CY) 1981).

The fitting of robust straight-lines by a variety of simple subgroup-oriented procedures has been compared by I. M. Johnstone and Paul Velleman (Velleman, with IMJ) 1985). A supplement analyzing their numerical results more thoroughly was prepared by John Tukey 1985o).

A practical approach to configural polysampling for straight-line regression has been pioneered by Fanny (Zambuto) O'Brien (Ph.D. Thesis 1984, Tech. Rep. 277, 278, 279, 282). This approach replaces 2-dimensional numerical quadrature by finite summation of algebraic expressions for a somewhat restricted family of joint-distribution shapes.

An approach by combined approximation and sampling methods to short-cutting the computational difficulties with Fanny O'Brien's approach is being investigated by Ha Nguyen (work in progress).

A comment on Huber's review of projection pursuit, including projection pursuit regression, was published by John Tukey (1985m).

*** strengthening theoretical understanding ***

Change-of-variance sensitivities in regression analysis have been studied by P. Rousseuw and Elvezio Ronchetti (Ronchetti (with PR) 1985).

A long-term program to attack the difficult questions of optimal fitting with 3 or more parameters, where other methods do not seem practical, was undertaken by Elvezio Ronchetti in 1983, using a small-sample-asymptotics approach. Early steps include a variational equation for bioptimal estimates of a single parameter (in progress).

Bounded-influence inference in regression has been studied by Elvezio Ronchetti (Tech. Rep. 257).

Robust C(α)-type tests for linear models have been studied by Elvezio Ronchetti (Tech. Rep. 258).

*** week ***

A regression week, where selected active workers in the field could interact effectively was held in Princeton 16-19 November 1983.

5d. Analysis-of-variance procedures (including factorialization and multiple comparisons)

The "analysis of variance", as applied to numbers that fall naturally into a factorial pattern, sometimes refers to the general idea of splitting up observed responses into (a common term,) main effects, and interactions of various orders, and sometimes to a specific way of reporting summary information about such splittings due to R. A. Fisher (more than 60 years ago). Alongside of regression procedures, analysis-of-variance procedures are certainly among those most frequently and most usefully applied to data. This section takes the more general interpretation and is divided into:

- topics within classical anova
- new value-splittings
- anova as it should be

The first of these comes closest to the narrower interpretation.

* topics within classical anova *

Higher-rank fits, going beyond the usual fitting of additive main effects, had been discussed by many authors, including D. R. McNeil and John Tukey (1975a). Applications of these fits to demography were studied by Mary Breckenridge (Tech. Rep. 143, Breckenridge 1983).

Robust analogs of the quantities arising in the classical analysis of variance were studied by Henry Braun and D. R. McNeil (1981).

Multiple comparison procedures seem to be inevitably needed as soon as main effects involving more than one degree of freedom, and not naturally divided into single degrees of freedom, are to be dealt with (e.g. Tukey 1949f). As the thirty-odd years have passed since such procedures first became prominent, a variety of procedures have been developed by several authors. In 1983, H. Braun and John Tukey (Tukey, (with HB) 1983c) reported on a new and promising procedure, specially adapted to the third (and perhaps the second) slice of a large family of such problems, where the slices can be roughly characterized by these verbal desires:

> 1st slice: I want to find at least one simultaneously significant comparison; I'll be glad to have more.
>
> 2nd slice: Surely this experiment measures well enough for some comparisons to be significant; I want to find as many significant as I can.
>
> 3rd slice: Surely this experiment measures well enough to find many significant comparisons; I want to find all that are really different.

In addition to these alternative desires for significance, there is a parallel desire for confidence, which should be answered quite differently.

* new value-splittings *

One of the innovations in *Exploratory Data Analysis* (Tukey 1977a) and its preliminary editions was the emphasis on median polish - - the taking out iteratively, in various directions, of medians - - as a way of splitting up the given values that appeared in a factorially-patterned table, whether row-by-column or more complicated. If all the fibers - - all the one-dimensional subarrays - - in such a table have zero medians, the table is said to have median balance. (A program for median polishing up to 7-way tables was prepared in 1980.)

By 1980, the advantages of modified medians - - when we have an even number of values to median, we have some choice along the interval joining the two central values - - began to be clear. (The *lomedian* is, in such a case the lower of the two central values.)

During 1980-83, Andrew Siegel studied the possibility of finding value-splitting with " many (exact) zeroes" - - splittings such that each subtable has no more non-zero entries than it has degrees of freedom. For the two-way table he found that many zeroes can always be had in combination with minimum L_1-norm and lomedian balance (Siegel 1983c). In 1985 Eugene Johnson (Ph.D. Thesis, forthcoming) showed that this is true for k-way tables with any k.

During 1982-83, Andrew Siegel studied the use of the scaled non-central chi-square distribution with zero degrees of freedom in modelling data sets containing exact zeroes (Tech. Rep. 239, Siegel 1985a).

* anova as it ought to be *

The original Fisher concept - - make the analysis-of-variance the experimental design called for, and stare at the table of mean squares - - took us a long, long way. In 1960, B. F. Green, Jr. and John Tukey reported on various extensions of this approach, as exemplified on a psychometric example of P. Johnson and F. Tsao. The implications of some of these extension did not all become clear for two decades. In particular, for a long while the introduction of an incomplete, well-tailored down-sweeping step to follow the upsweeping corresponding to the standard fitting was not recognized as a general characteristic to be sought in both robust/resistant and more conventional analysis of variance.

In 1980-82, Allan Seheult and John Tukey (Tukey (with AS) 1982i) updated and expounded a robust/resistant analysis for 2^n patterns of data (n factors each in 2 version or at 2 level).

At about the same time, Allan Seheult and John Tukey recognized that a robust version of the upsweeping process could, and plausibly should, involve (i) a many-zero fitting, (ii) identification and setting aside or modification of idiosyncratic entries in each subtable (sometimes called "assassination") and (iii) least-square fitting to the resulting tables. (Unpublished work at Bell Laboratories by E. Fowles and J. McRae made useful contributions here.) (Though still unpublished, this work had important influences on the three items to be next reported.)

In 1984-86, Eugene Johnson studied carefully the behavior of such a three-stage procedure. He found a Gaussian-slash bi-efficiency of 95% achievable by such a procedure (forthcoming Ph.D. Thesis). By comparison with bi-efficiencies of centering single samples, this seems to be a very high bi-efficiency indeed.

In 1985, Eugene Johnson and John Tukey (1986***) prepared an account, based in part on Johnson's tools developed in his thesis, of a graphical, dissect-everything-into-single-degrees-of-freedom approach to a classical analysis of variance that seemed to deserve the title "Graphical Exploratory Analysis of Variance". (The robust analog will be presented in Eugene Johnson's forthcoming Ph.D.) Thesis.

Starting in 1985, John Tukey is developing successive drafts of a long paper on what a good computer data-analysis system ought to do. One long section deals with approaches to factorialization (work in progress).

5e. Spectrum analysis procedures

Modern spectrum analysis is thirty-odd years old (cp Tukey 1984b), and has seen many steps forward.

The problem of assessing the underlying power spectrum of a time series when our observations are contaminated by "spiky noise" was studied by Michael Schwarzchild (Ph.D. Thesis 1975) with encouraging results.

A discussion of the "styles" of spectrum analysis was prepared and published by John Tukey 1984a).

The relationship between the two most important of these styles: classical spectrum analysis - - the method of choice for Gaussianly distributed time series - - and maximum entropy spectrum analysis (and related procedures) - - the method of choice for some very non-Gaussian time series - - was studied, with interesting and helpful results, by Clifford Hurvich (Ph.D. Thesis 1985, Hurvich 1985, 1986).

5f. Randomization in experimentation

The use of randomization in the assignment of individuals (plots, runs, people) to circumstances (treatments, conditions, etc.), coupled with an analysis of the results through the use of

i) simple, mathematically easily manipulable summaries

ii) randomization moments for these summary statistics obtained by formula-manipulation mathematics, and

iii) a hoped-for (and often obtained approach to normality of distribution for those statistics

is classical (much is about a half century old). This approach severely limited the summaries that could be used - - ruling out much of what robust/resistant techniques had to offer - - and failed to take advantage of modern computing capabilities.

The introduction of a way of using, and analyzing the results of randomization that was compatible with modern computing - - a way in which a balanced subset of all possible randomizations is chosen, the data is analyzed as if each had been used, and the position, among all these results, of the value of the key statistic for the actual randomization used is employed - - to give a significance test available for any choice of key statistics - - seems to be due to D. R. Brillinger, L. W. Jones and John Tukey (*Tukey (with DRB, and LWJ) 1978g) (work for the Weather Modification Advisory Board of the Department of Commerce). Under ARO sponsorship, John Tukey has carried this work considerably further (Tukey 1985l, Tukey 1986****).

5g. Other procedures

A somewhat diverse group of topics belong here, falling naturally into a few categories.

* goodness of fit and modified chi-square *

During 1977-78 Henry Braun studied the question of testing for goodness-of-fit in the presence of nuisance parameters (Braun 1980a).

A modified chi-square, stemming from earlier work by Freeman and Tukey (*Tukey (with MFF) 1950h) was suggested for use in Mosteller and Tukey (with FM) 1977b). Later work by John Tukey (in 1978-79) showed that this modification gave much closer approximations to the moments of tabular chi-square then the classical form when a Poisson approximation was appropriate (whenever the total sample size can be thought of as at least as variable as a poisson quantity). Later work by K. Larntz (*Larntz 1978) showed that in those goodness-of-fit problems where there are many cells of equal expected size, and the total sample was fixed, there are discomforts about the tail probabilities. An improved modification has been proposed, and its study begun (by Roger Pinkham and John Tukey).

* nominal-by-ordinal contingency tables *

A technical report on this subject by Gary Simon was revised and issued in 1974 (Tech.

Rep. 32, Simon 1974).

*** jackknifing discontinuous or nearly discontinuous statistics ***

John Tukey has recently developed a multiple split-half jackknife which promises good behavior for very uncomfortable cases (unpublished Statistics 411 notes).

*** consistent estimation in certain stochastic processes ***

P. Sampson and Andrew Siegel ((with PS) 1985b) have investigated consistent estimation in partially observed random walks.

*** combination of results ***

Frederick Mosteller and John Tukey have spent considerable effort on the bundle of problems involved in the many kinds of combination of results. Two papers (Tukey (with FM) 1982a and 1982e) have been published, and considerable material directed toward a book has been accumulated.

5h. Supporting technology

*** random search in global optimization ***

Work on this topic by Peter Bloomfield (with RSA) was reported in 1976 (Bloomfield 1976).

*** random integration methods unbiased for low-order polynomials**

Andrew Siegel and Fanny (Zambuto) O'Brien have developed random few-point designs for integration over rectangles that exact for low-degree polynomials and unbiased for any integral function (Tech. Rep. 226, Siegel (with FZ) 1983a, Siegel (with FZO) 1984a).

*** approximations to standard distributions ***

Starting from his co-editorial responsibility for 1985b, Anita Parunak and John Tukey have been working on the development of a satisfactory continuous approximation to the tail area of the hypergeometric distribution. (This has a convenience-only was in avoiding calculation of factorials for possibly large numbers, and a non-replaceable use as a continuous-parameter distribution with which to approximate various discrete distributions.)

John Tukey is part way through the development of a good continuous approximation. (same comment) to the F-distribution

*** antithetic variates in experimental sampling ***

Andrew Siegel has found a method of this sort (unpublished).

*** orthogonal arrays in experimental sampling ***

The use of orthogonal arrays as a source of balance and thus of improved variance in experimental sampling has been studied by Dhammika Amaratunga (Thesis, 1984).

5i. Probability questions

While diverse, many of these problems were suggested by statistical or data-analytic questions.

*** times to extinction for branching processes ***

Method for bounding the distributions of these times were sought by Henry Braun, and reported (Tech. Rep. 106, Braun 1975 and 1978) as "Polynomial bounds for probability generating functions".

*** coverage problems on the circle ***

Results by Andrew Siegel and Lars Holst have been reported (Siegel and Holst 1982d).

*** breakage problems, with application to species abundance ***

Results by Andrew Siegel and G. Sugihara have been reported (Tech. Rep. 215, Siegel and Sugihara 1983b).

* distances between all pairs of points of a random set *

Andrew Siegel has found an exact integral representation for the distribution of these distances. Various asymptotic results are then easily obtained (unpublished).

5j. Other topics

This is the catchall.

* miscellaneous applications *

In 1977, John Tukey talked on what modern statistical techniques might do for forecasting (unpublished).

In 1981, A. J. Arnold and Andrew Siegel studied the regularity of triple points of tectonic plates on the earth's surface. (Tech. Rep. 216).

PROGRESS REPORT

(TWENTY COPIES REQUIRED)

1. ARO PROPCSAL NUMBER: 19442-MA

2. PERIOD COVERED BY REPORT: July 1, 1985 - January 31, 1986

3. TITLE OF PROPOSAL: More Realistic Techniques for Data Analysis

4. CONTRACT OR GRANT NUMBER: DAAG29-82-K-0178

5. NAME OF INSTITUTION: Princeton University

6. AUTHOR(s) OF REPORT: John W. Tukey

7. LIST OF MANUSCRIPTS SUBMITTED OR PUBLISHED UNDER ARO SPONSORSHIP DURING THIS PERIOD, INCLUDING JOURNAL REFERENCES:

(a) Hoaglin, David C., Mosteller, Frederick, Tukey, John W. (eds.) (1985). Exploring Data Tables, Trends, and Shapes, John Wiley & Sons, Inc. Publishers, New York.

(b) Hampel, Frank R., Ronchetti, Elvezio M., Rousseeuw, Peter J., Stahel, Werner A., (eds.) (1986). Robust Statistics: The Approach Based on Influence Functions, John Wiley & Sons, Inc. Publishers, New York.

(c) Hoaglin, David C., Tukey, John W. (1985). "Checking the discrete distributions," Exploring Data Tables, Trends, and Shapes, John Wiley & Sons, Inc. Publishers, New York, Chapter 9: 345-416.

8. SCIENTIFIC PERSONNEL SUPPORTED BY THIS PROJECT AND DEGREES AWARDED DURING THIS REPORTING PERIOD:

Graduate Research Assistant:
G. Easton, Ph.D. awarded August 1985
E. Johnson, 100% July through November 1
H. Nguyen, 100% July 1985 through January 31, 1986
Research Assistant:
E. Olszewski, 100% July, August; 50% September through January 31, 1986
Assistant Professor:
Elvezio Ronchetti, 100% July, August
Colin Goodall, 100% July, August
Principal Investigator:
John W. Tukey, 20% July through October
20% to 0% November, December 1985 and January 1986

John W. Tukey 19442-MA
PRINCETON UNIVERSITY
DEPARTMENT OF STATISTICS
PRINCETON NJ 08544

7. List of Manuscripts (cont'd)

(d) Tukey, John. W. (1985). "Improving crucial randomized experiments - - especially in weather modification - - by double randomization and rank combination," *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. 1, Lucien M. Le Cam and Richard A. Olshen, eds., 330-359.

(e) Tukey, John W. (1985) "Discussion" of Peter Huber's paper on Projection pursuit, Annals of Statistics, Vol. 13, No. 2, 517-518, June 1985.

(f) Tukey, John W. (1985) "The variance of slopes of lines fitted groups: an analysis of the Johnstone and Velleman Monte Carlo results," *J. Amer. Statist. Assoc.* 80: 1055-1059.

(g) Tukey, John W. (1986) "Sunset Salvo," (based on material presented or handed out at the 1985 Spring Symposium of the Northern New Jersey Chapter of ASA, 6 May 1985), *Vol. 40, No. 1*, 72-76.

(h) Tukey, John W. (1986) "Limited randomization with detailed reassignment as the key to taking advantage of modern summaries," to appear in *Proceedings of the 30th Conference on the Design of Experiements in Army Research, Development and Testing*, Army Research Office, Durham, North Carolina.

(i) Tukey, John W. (1986) "Configural polysampling," (presented as The John von Neumann Lecture for 1985 at the SIAM national meeting in Pittsburgh, June 24-26, 1985), to appear in the SIAM Review.

(j) Johnson, Eugene G., Tukey, John W. (1986) "Graphical exploratory analysis of variance illustrated on a splitting of the Johnson and Tsao data," (to appear in a book in honor of Cuthbert Daniels's 80th birthday), to be published by Wiley & Sons, Inc. Publishers, New York.

BRIEF OUTLINE OF RESEARCH FINDINGS

1. Ongoing research

Small sample asymptotics for regression. Bioptimal estimators for regression are available, but their computation pose serious problems in more than two dimensions. One way to cope with this problem is to restrict the search for the best estimator in the class of M-estimators.

By means of small sample asymptotics techniques, one can approximate the mean square error of an M-estimator under two (or more) sampling situations. Then one can write the variational equation that the bioptimal or polyoptimal (in the sense that it cannot be improved in all sampling situations simultaneously) M-estimator must satisfy. This equation has to be solved numerically.

Work has concentrated on the problem of finding the optimal M-estimator for location models under a single situation. The final goal of this project is to derive bioptimal M-estimators for regression. (E. Ronchetti)

Robustness in nonlinear models. The goal of this project is to derive robust procedures for some nonlinear models. We begin with a simple problem.

Let F(x,y) be a bivariate distribution, spherically symmetric, with center of symmetry at the origin. The data consists of n iid observations $(x_1, y_1), \ldots, (y_n, y_n)$ with distribution $F(x - \rho\cos\theta, y - \rho\sin\theta)$. The parameter of interest is θ. The constant ρ is assumed known and determines the degree of nonlinearity in the problem.

The inference problem described is invariant under rotations. First we can derive the best equivariant estimator under a single situation. Secondly, by applying the theory developed for location and linear regression models, it is possible to compute bioptimal and polyoptimal estimators which cannot be improved in all sampling situations simultaneously. (E. Ronchetti, S. Morgenthaler, Dept. of Statist., Yale University)

Aspects of the analysis of data from factorial designs. One set of work was with John Tukey on graphical techniques for exploratory analysis of factorial data sets. These techniques are extensions of the half-normal plots of Daniel and compared the ordered absolute values of the normalized single degree of freedom contrasts with typical values of order statistics from the half-Gaussian distribution. Among other things, these techniques allow the selection of apparently important contrasts, can indicate the need to reformulate the data, and can indicate the need to select a different set of defining contrasts. Details are given in the paper: "Graphical exploratory analysis of variance". (E. Johnson, J. W. Tukey)

Doctoral research of Eugene Johnson has been concerned with the robust analysis of data from complete, unreplicated, factorial designs. My approach consists of a three-stage procedure. The first stage consists of obtaining a resistant fit to the table where there are exactly as many nonzero residuals as there are residual (highest order interaction) degrees of freedom. Such a fit is called an elemental fit and the set of p(= rank of the design matrix) observations corresponding to the zero residuals is called an elemental subset.

An important issue is the choice of the particular fitting procedure used to obtain the elemental fit. If all goes well, the effects of any exotic observations will be confined to the nonzero residuals. The next phase consists of screening the nonzero residuals from the elemental fit for potential outliers by comparing the magnitudes of the ordered absolute residuals with reference values from a half-Gaussian distribution. The third phase consists of cleaning the data by removing any declared outliers, replacing them with missing values estimates and then conducting a least-squares analysis of the result with appropriate adjustment of the constituent degrees of freedom. A major determinant of the ultimate performance of the entire three-phase technique is the breakdown characteristics of the phase I fitting technique used.

A successful approach is to use a variant of median polish in which, at each step, a median is swept from the set of values which are candidates for estimating a given parameter. By proper choice of the order of fitting and of the type of median to be used, a high overall breakdown point can be achieved. Experimental sampling results indicate that a Gaussian-slash biefficiency of 85% is achievable by applying the second and third phases to the results of such an elemental fit. (E. Johnson)

Retabing distributional systems. The 4- and 6- parameter g -- and h — normal-transformation families (Hoaglin, 1985) offer a considerable range of distributional shapes. A particular advantage is that a distribution can be fit *resistantly*, directly to the data quantiles, avoiding the use of highly variable 3rd and 4th sample moments to fit a non-normal shape. How close are the g — and h — distributions to members of classical families, in particular the Pearson family? Moment-matching (Martinez & Iglewicz, 1984) is poor for percentiles. This is a serious flaw, in light of the resistant quantile-fitting philosophy underlying the g — and h — families. Calculations of Jim Landwehr match percentiles to provide two g — and two h — parameters to 5 S.F. for arrays of $(\sqrt{\beta_1}, \beta_2)$ values among the Pearson Type IV, VI & VII distributions.

The present effort uses Landwehr's data to provide parsimonious algebraic expressions mapping (β_1, β_2) to the g — and h — parameters. Some intermediate results, e.g. of the form

$$g_o = a_{g_o}\sqrt{\beta_1}\left\{1 + \frac{(b_{g_o} + c_{g_o}\beta_1)}{\beta_2^2}\right\}$$

are accurate to 3-4 S.F. Better precision is expected. The goal, an invertible transformation (or pair of transformations)

$$\left\{u, \sigma^2, \beta_1, \beta_2\right\} \longleftrightarrow \left\{u, \sigma^2, g_o, h_o, (g_2, h_2)\right\}$$

will be acceptable if percentage points *and* parameters (moment-based and reshaping - g's and h's) match well.

The algebraic forms of the transformations are unknown. involves generalizations of existing techniques for two-way tables (Emerson & Wong, 1985): triangular non-additive tables of g — and h — parameters (indexed by $\sqrt{\beta_1}$ and β_2) are transformed to square additive tables. Iterative proportional fitting of a generalized additive model, an approach outlined by J.W. Tukey, is being developed to provide an invertible mapping $R^2 \rightarrow R^2$. *Given* a functional form, the function parameters may be optimized directly. The ACE algorithm is a related technique for comparison. (Colin Goodall)

Statistics of change in size and shape for landmark data. Consider a sample of biological organisms, or other geometrical forms, each described by the same set of homologous point landmarks. The landmark co-ordinates are measured at two different times, in between which growth may have occurred. Statistical analysis of change in size of each form is based on a univariate summary, e.g. ratio of mean-square dispersion about the centroid or ratio of areas. Statistical analysis of change in shape involves a multivariate statistic, essentially the residuals from the a fit of translation, isotropic scale, and rotation (similarity transformation) to each pair of forms. Multivariate tests, for example the one- and two-sample location problems for mean shape change, may be based on this statistic. In Goodall (1986) I propose, in general terms, a model comprising a continuously-varying deformation tensor field with explicit components for measurement error, shape change, and inter-individual variation. I also elucidate the connection between the least-squares fit of an affine transformation to a triangle of landmarks and Bookstein's geometrical method *(op cit)*. In later work I develop a classical multivariate approach to the analysis of shape change in samples of pairs of forms based on Procrustes methods for least-squares fitting. Statistical properties are derived via perturbation analysis. (Colin Goodall)

Engineering methods applied to models of plant growth. Ongoing research at Stanford University suggests that morphological development in plants may be mediated in an essential way by the interaction, in the epidermis, of turgor pressure, global geometry, and cell-specific directional reinforcement by cellulose microfibrils (Green and Poethig, 1984). Cell polarity (reinforcement direction), cell division direction, and cell deformation appear closely interrelated by genetically-determined "rules" of cell activity. A finite element model of plant growth has such rules embedded in its constitutive equations (Goodall, 1985). Specific details include an abstraction of cell geometry and a scheme for simulations based on alternative rule sets. (Colin Goodall)

Smoothing. A number of drafts of a paper on "Thinking about smoothing" have been written, and a Technical Report will soon be prepared. (J. W. Tukey)

Clustering/dissecting. A number of generations of improvement have been conducted for procedures for dissecting a set of points (usually in the plane) consisting of a mixture of 3 samples of 50; one from each of three spherical Gaussian distributions centered at the vertices of an equilateral triangle. Performance, given only the 150 points (without distriction between samples) and that 3 pieces are to make, at mutual separations of 3.7σ is extremely close to that which knowledge of the population will permit; at 3.2σ it is still close. The problem of learning from the data how many pieces to seek has not been addressed. (Katherine Hansen, J. W. Tukey)

Interactive analysis. Motivated by the need to display effectively what has already been tried on a data set, John Tukey has prepared several drafts of an account of what operations a relatively complete computing system for data analysis might provide, and how the history of the analysis might be summarized. (J. W. Tukey)

END

DTIC

7-86